

Claims

What is claimed is:

1 *Sub A1* 1. A method for identifying groups of pages of common interest from a
2 collection of hyper-linked pages, comprising the steps of:
3 identifying a plurality of community cores from the collection, each core being
4 first and second sets of pages, each page in the first set pointing to every page in
5 the second set; and
6 expanding each identified core into a full community, the full community
7 being a subset of the pages regarding a particular topic.

1 2. The method as recited in claim 1, wherein:
2 the collection includes a plurality of sites, each site having one or more
3 hyper-linked pages; and
4 the method further includes the step of removing the hyper-links between
5 any two pages on a same site.

1 *Sub A2* 3. The method as recited in claim 2 further comprising the step of
2 discarding the pages of predetermined sites.

1 4. The method as recited in claim 1 further comprising the steps of:
2 finding highly similar pages that have different names;
3 replacing the highly similar pages with a single representative page; and

4 redirecting any hyper-links that pointed to one of the highly similar pages so
5 that the redirected hyper-links now point to the representative page.

1 5. The method as recited in claim 1 further comprising the steps of:
2 discarding unnecessary pages from consideration to generate a set of
3 candidate fan pages and a set of candidate center pages; and
4 using the set of candidate fan pages and set of candidate center pages as
5 the first and second sets, respectively, to identify the community cores.

1 6. The method as recited in claim 5, wherein the step of discarding
2 includes the steps of:
3 determining candidate fan pages, the candidate fan pages being those
4 pointing to at least a predetermined number of different sites;
5 determining candidate center pages, the candidate center pages being those
6 pointed to by one or more candidate fan pages; and
7 discarding all pages in the collection except the candidate fan pages and
8 candidate center pages.

1 7. The method as recited in claim 6, wherein the determination of
2 candidate fan pages is based on page content and the hyper-links pointing
3 therefrom.

1 8. The method as recited in claim 5, wherein the step of identifying a
2 plurality of community cores includes the step of finding a plurality of (i, j)-cores
3 where i and j are the numbers of candidate fan pages and candidate center pages,
4 respectively, that appear in each identified community core.

1 9. The method as recited in claim 8, wherein the step of finding a
2 plurality of (i, j)-cores includes the steps of:

3 (a) discarding all candidate center pages that have fewer than i hyper-links
4 pointing thereto;
5 (b) determining all candidate center pages that have i hyper-links pointing
6 thereto and determining whether the i hyper-links represent a valid community core;
7 and
8 (c) if the i hyper-links represent a valid community core, then outputting the
9 valid core, otherwise, discarding the determined candidate center pages.

1 10. The method as recited in claim 9 further comprising the steps of:

2 (d) discarding all candidate fan pages that have fewer than j hyper-links
3 pointing therefrom;
4 (e) determining all candidate fan pages that have j hyper-links pointing
5 therefrom and determining whether the j hyper-links represent a valid community
6 core; and
7 (f) if the j hyper-links represent a valid community core, then outputting the
8 valid core, otherwise, discarding the determined candidate fan pages.

1 11. The method as recited in claim 10 further comprising the step of
2 repeating steps (a)-(f) until every candidate fan page has more than j hyper-links
3 pointing therefrom and every candidate center page has more than i hyper-links
4 pointing thereto.

1 12. The method as recited in claim 10 further comprising the step of
2 repeating steps (a)-(f) until a predetermined ending condition is satisfied.

1 13. The method as recited in claim 10 further comprising the steps of:
2 determining all $(2,j)$ cores by examining all pairs of candidate fan pages;
3 for $i = 3$ to n , where n is a predetermined value:
4 (i) finding all (i,j) -cores by examining the $(i-1,j)$ -cores; and
5 (ii) for each $(i-1, j)$ -core, determining whether any of the candidate fan
6 pages may be added to the $(i-1, j)$ -core to yield a (i,j) -core; and
7 removing all (i,j) -cores that appear as subsets of (i',j) cores, where $i' > i$.

1 14. A computer program product for use with a computer system for
2 identifying groups of pages of common interest from a collection of hyper-linked
3 pages, the computer program product comprising:
4 a computer-readable medium;
5 means, provided on the computer-readable medium, for directing the system
6 to identify a plurality of community cores from the collection, each core being first

7 and second sets of pages, each page in the first set pointing to every page in the
8 second set; and

9 means, provided on the computer-readable medium, for directing the system
10 to expand each identified core into a full community, the full community being a
11 subset of the pages regarding a particular topic.

1 15. The computer program product as recited in claim 14, wherein:
2 the collection includes a plurality of sites, each site having one or more
3 hyper-linked pages; and

4 the product further includes means, provided on the computer-readable
5 medium, for directing the system to remove the hyper-links between any two pages
6 on a same site.

1 16. The computer program product as recited in claim 15 further
2 comprising means, provided on the computer-readable medium, for directing the
3 system to discard the pages of predetermined sites.

1 17. The computer program product as recited in claim 14 further
2 comprising:

3 means, provided on the computer-readable medium, for directing the system
4 to find highly similar pages that have different names;

5 means, provided on the computer-readable medium, for directing the system
6 to replace the highly similar pages with a single representative page; and

7 means, provided on the computer-readable medium, for directing the system
8 to redirect any hyper-links that pointed to one of the highly similar pages so that the
9 redirected hyper-links now point to the representative page.

1 18. The computer program product as recited in claim 14 further
2 comprising:

3 means, provided on the computer-readable medium, for directing the system
4 to discard unnecessary pages from consideration to generate a set of candidate fan
5 pages and a set of candidate center pages; and

6 means, provided on the computer-readable medium, for directing the system
7 to use the set of candidate fan pages and set of candidate center pages as the first
8 and second sets, respectively, to identify the community cores.

1 19. The computer program product as recited in claim 18, wherein the
2 means for directing to discard includes:

3 means, provided on the computer-readable medium, for directing the system
4 to determine candidate fan pages, the candidate fan pages being those pointing to
5 at least a predetermined number of different sites;

6 means, provided on the computer-readable medium, for directing the system
7 to determine candidate center pages, the candidate center pages being those
8 pointed to by one or more candidate fan pages; and

9 means, provided on the computer-readable medium, for directing the system
10 to discard all pages in the collection except the candidate fan pages and candidate
11 center pages.

1 20. The computer program product as recited in claim 19, wherein the
2 determination of candidate fan pages is based on page content and the hyper-links
3 pointing therefrom.

1 21. The computer program product as recited in claim 18, the means for
2 directing to identify a plurality of community cores includes means, provided on the
3 computer-readable medium, for directing the system to find a plurality of (i, j)-cores
4 where i and j are the numbers of candidate fan pages and candidate center pages,
5 respectively, that appear in each identified community core.

1 22. The computer program product as recited in claim 21, wherein the
2 means for directing to find a plurality of (i, j)-cores includes:

3 (a) means, provided on the computer-readable medium, for directing the
4 system to discard all candidate center pages that have fewer than i hyper-links
5 pointing thereto;
6 (b) means, provided on the computer-readable medium, for directing the
7 system to determine all candidate center pages that have i hyper-links pointing
8 thereto and determining whether the i hyper-links represent a valid community core;
9 and

10 (c) means, provided on the computer-readable medium, for directing the
11 system to output the valid core if the i hyper-links represent a valid community core,
12 otherwise, to discard the determined candidate center pages.

1 23. The computer program product as recited in claim 22 further
2 comprising:

3 (d) means, provided on the computer-readable medium, for directing the
4 system to discard all candidate fan pages that have fewer than j hyper-links pointing
5 therefrom;

6 (e) means, provided on the computer-readable medium, for directing the
7 system to determine all candidate fan pages that have j hyper-links pointing
8 therefrom and determining whether the j hyper-links represent a valid community
9 core; and

10 (f) means, provided on the computer-readable medium, for directing the
11 system to output the valid core if the j hyper-links represent a valid community core,
12 otherwise, discard the determined candidate fan pages.

1 24. The computer program product as recited in claim 23, wherein the
2 operation of means (a)-(f) is repeated until every candidate fan page has more than
3 j hyper-links pointing therefrom and every candidate center page has more than i
4 hyper-links pointing thereto.

1 25. The computer program product as recited in claim 23, wherein the
2 operation of means (a)-(f) is repeated until a predetermined ending condition is
3 satisfied.

1 26. The computer program product as recited in claim 23 further
2 comprising:

3 means, provided on the computer-readable medium, for directing the system
4 to determine all (2,j) cores by examining all pairs of candidate fan pages;

5 for $i = 3$ to n , where n is a predetermined value:

6 (i) means, provided on the computer-readable medium, for directing
7 the system to find all (i,j)-cores by examining the (i-1,j)-cores; and

8 (ii) for each (i-1, j)-core, means, provided on the computer-readable
9 medium, for directing the system to determine whether any of the candidate fan
10 pages may be added to the (i-1, j)-core to yield a (i,j)-core; and

11 means, provided on the computer-readable medium, for directing the system
12 to remove all (i,j)-cores that appear as subsets of (i',j) cores, where $i' > i$.

1 27. A system for identifying groups of pages of common interest from a
2 collection of hyper-linked pages, comprising:

3 means for identifying a plurality of community cores from the collection, each
4 core being first and second sets of pages, each page in the first set pointing to
5 every page in the second set; and

6 means for expanding each identified core into a full community, the full
7 community being a subset of the pages regarding a particular topic.

1 28. The system as recited in claim 27, wherein:

2 the collection includes a plurality of sites, each site having one or more
3 hyper-linked pages; and

4 the method further includes the step of removing the hyper-links between
5 any two pages on a same site.

1 29. The system as recited in claim 28 further comprising means for
2 discarding the pages of predetermined sites.

1 30. The system as recited in claim 27 further comprising:
2 means for finding highly similar pages that have different names;
3 means for replacing the highly similar pages with a single representative
4 page; and
5 means for redirecting any hyper-links that pointed to one of the highly similar
6 pages so that the redirected hyper-links now point to the representative page.

1 31. The system as recited in claim 27 further comprising:
2 means for discarding unnecessary pages from consideration to generate a
3 set of candidate fan pages and a set of candidate center pages; and
4 means for using the set of candidate fan pages and set of candidate center
5 pages as the first and second sets, respectively, to identify the community cores.

1 32. The system as recited in claim 31, wherein the means for discarding
2 includes:

3 means for determining candidate fan pages, the candidate fan pages being
4 those pointing to at least a predetermined number of different sites;

5 means for determining candidate center pages, the candidate center pages
6 being those pointed to by one or more candidate fan pages; and

7 means for discarding all pages in the collection except the candidate fan
8 pages and candidate center pages.

1 33. The system as recited in claim 32, wherein the determination of
2 candidate fan pages is based on page content and the hyper-links pointing
3 therefrom.

1 34. The system as recited in claim 31, the means for identifying a plurality
2 of community cores includes means for finding a plurality of (i, j) -cores where i and j
3 are the numbers of candidate fan pages and candidate center pages, respectively,
4 that appear in each identified community core.

1 35. The system as recited in claim 34, wherein the means for finding a
2 plurality of (i, j) -cores includes:

3 (a) means for discarding all candidate center pages that have fewer than i
4 hyper-links pointing thereto;

5 (b) means for determining all candidate center pages that have i hyper-links
6 pointing thereto and determining whether the i hyper-links represent a valid
7 community core; and
8 (c) means for outputting the valid core if the i hyper-links represent a valid
9 community core, otherwise, discarding the determined candidate center pages.

1 36. The system as recited in claim 35 further comprising:

2 (d) means for discarding all candidate fan pages that have fewer than j
3 hyper-links pointing therefrom;
4 (e) means for determining all candidate fan pages that have j hyper-links
5 pointing therefrom and determining whether the j hyper-links represent a valid
6 community core; and
7 (f) means for outputting the valid core if the j hyper-links represent a valid
8 community core, otherwise, discarding the determined candidate fan pages.

1 37. The system as recited in claim 36, wherein the operation of means

2 (a)-(f) is repeated until every candidate fan page has more than j hyper-links
3 pointing therefrom and every candidate center page has more than i hyper-links
4 pointing thereto.

1 38. The system as recited in claim 36, wherein the operation of means

2 (a)-(f) is repeated until a predetermined ending condition is satisfied.

1 39. The system as recited in claim 36 further comprising:
2 means for determining all (2,j) cores by examining all pairs of candidate fan
3 pages;
4 for $i = 3$ to n , where n is a predetermined value:
5 (i) means for finding all (i,j) -cores by examining the $(i-1,j)$ -cores; and
6 (ii) for each $(i-1, j)$ -core, means for determining whether any of the
7 candidate fan pages may be added to the $(i-1, j)$ -core to yield a (i,j) -core; and
8 means for removing all (i,j) -cores that appear as subsets of (i',j) cores, where
9 $i' > i$.